

Computer vision that can ‘see’ in the dark

Shi Yong Goh¹, Yan Chiew Wong¹, Syafeeza Ahmad Radzi¹, Ranjit Singh Sarban Singh²

¹Faculty of Electronics and Computer Technology and Engineering, Universiti Teknikal Malaysia Melaka (UTeM), Melaka, Malaysia

²School of Engineering and Technology, Sunway University, Selangor, Malaysia

Article Info

Article history:

Received Sep 6, 2023

Revised Feb 4, 2024

Accepted Mar 1, 2024

Keywords:

Computer vision

Convolutional neural network

Dark frame enhancement

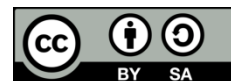
Human action recognition

YOLOv7

ABSTRACT

Insufficient lighting environment has raised challenges for night shift workers' safety monitoring. Thus, we have developed a computer vision-based algorithm recognizing 11 actions based on action recognition in dark (ARID) dataset. A hybrid model of integrating convolutional neural network (CNN) into YOLOv7 has been proposed. YOLOv7 is an algorithm designed for real-time object detection in image or video, for fast and accurate detection in applications such as autonomous vehicles and surveillance systems. In this work, video in dark environment has first been enhanced using CNN algorithm before feeding into YOLOv7 network for activity recognition. Adaptive gamma intensity correction (GIC) has been integrated to further improving the overall result. The proposed model has been evaluated over different enhancement modes. The proposed model is able to handle dark video frames with 74.95% Top-1 accuracy with fast processing speed of 93.99 ms/frame on a 4 GB RTX 3050 graphical processing unit (GPU) and 17.59 ms/frame on 16 GB Tesla T4 GPU. The base size of the proposed model is tiny, only 74.8 MB, but with 36.54 M of total parameters indicating that it has more capacity to learn more meaningful information with limited hardware resources.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Wong Yan Chiew

Faculty of Electronics and Computer Technology and Engineering, Universiti Teknikal Malaysia Melaka (UTeM)
Durian Tunggal, Melaka 76100, Malaysia

Email: ycwong@utem.edu.my

1. INTRODUCTION

Video surveillance in dark environments is still challenging as the dark scenes taken are degraded due to noises caused by insufficient lighting condition. It often requires external hardware for night surveillance and hence increases the cost. Most algorithms only focus on detection to detect anomalies and gather visual evidence [1], but not consider the object's action, hence is limited for human safety monitoring in dark incident scenes such as workplace. In recent years, human action recognition (HAR) had become a significant attention in the field of computer vision (CV) [2]. HAR in dark videos involves several steps start from collecting raw data until the conclusion about the desired action, those processes include data preprocessing, feature extraction, and action recognition by using CV algorithm [3]. The only difference is that dark condition HAR requires additional step to called image enhancement. HAR in dark can be achieved with the help of image enhancement technique to make dark image more visually receptive. A straightforward approach is to preprocess the dark image manually by applying classical methods such as histogram equalization (HE) and gamma intensity correction (GIC). To have a more adaptive and effective enhancement process, the task relies on advanced methods that utilize neural networks.

Various neural network-based enhancement techniques have been invented including exposure [4] that utilised generative adversarial networks (GAN) network with deep reinforcement learning approach to

learn decision-making on action to be taken to facilitate the operation of a set of filters for image enhancement, iterative convolutional neural network (CNN) [5] that use fully convolutional network (FCN) network to learn high dynamic range (HDR) image's feature to enhance the low dynamic range (LDR) dark image, seeing motion in the dark (SID) [1] that use two-pathway framework that use pairs of low-light video and corresponding long-exposure version for enhancement by linear scale dark frame pixel to match the brightness and dynamic range of the long-exposure frame, and kindling the darkness (KinD) [6] that deals with extraction of illumination information from input images and perform adjustment and use a flexible mapping function learned from real data to adjust light level in an image.

HAR often involves the extraction of spatial and temporal features of a given action and analyzing those features to perform action recognition tasks. However, huge number of features leads to high computational resources. Object detection algorithms, originally used for identifying and localizing objects of interest within an image, are useful and more computationally friendly for recognizing actions. One-stage regression-based approaches such as the YOLO series [7], [8] had been utilized in [9] for such purposes by providing labels and bounding box coordinates to the person performing action in each video frame and predicting it by a single CNN.

Current approaches separate action recognition and dark image enhancement into two separated stages [10]. A recently launched model uses a combination of domain adaptable normalization (DANorm) and R(2+1)D-34 [11] to focus on features normalization, angle constraint for preventing misclassification between labelled and unlabelled dataset, and Pseudo-label (method to solve data imbalance issue). Other methods such as combination of Zero-reference deep curve estimation (Zero-DCE) as image enhancer and R(2+1)D as action recognizer [12], Dark-Light + R(2+1)D-34 [13] are also being considered as the references. In this work, a two-stage approach consisting of dark frame enhancement (DFE) model and YOLOv7 detector as an action recognizer for HAR in dark videos is proposed. The effectiveness of DFE to facilitate the YOLOv7 for HAR was investigated in terms of mAP@0.5 and single frame inference speed.

2. METHOD

The pipeline shown in Figure 1 consists of a CNN-based enhancement model (DFE) and an action recognizer (YOLOv7) that trained in a supervised manner. The input video frame was first resized into 256×256 size and fed into DFE for enhancement. The frame filtered by DFE is treated as the input for YOLOv7. The DFE is further integrated with adaptive GIC technique to select a suitable value for a parameter called gamma in an automated way for getting optimal enhancement in image's brightness and contrast.

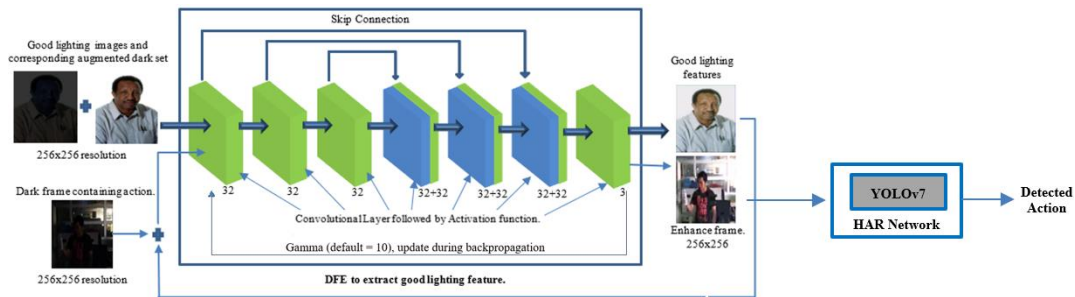


Figure 1. The training pipeline of the proposed DFE+YOLOv7 model

2.1. Convolutional neural network-based dark frame enhancement model

The model architecture has been designed to be resolution independent. The DFE has only several CNN layers stacked together to extract lighting features from well-lit images and to learn a good mapping of the dark video frames toward the good lighting features. There are skip connections within the stacked CNN layers to improve information flow and reduce the vanishing gradient problem by learning both low-level and high-level features. The proposed DFE model contains only 67, 299 k parameters.

The proposed adaptive GIC technique first applies GIC on each frame pixel based on (1). The value of gamma parameter, γ will affect the frame's brightness and contrast such that, $\gamma = 1.0$ will have no changes on the brightness and contrast, $\gamma > 1.0$ will increase both and vice versa. γ is differentiable and has been introduced into the loss function based on (2) to further improve the model performance.

$$L_{V(out)} = L_{V(in)}^\gamma \quad (1)$$

where $L_{V(out)}$ is enhanced frame pixel luminance value, $L_{V(in)}$ is original frame pixel luminance value, and γ is gamma value.

$$\begin{aligned} loss(mse) &= \frac{1}{n} \sum (\log_{10}(L_{V(in)}) - \log_{10}(L_{V(out)}))^2 \\ &= \frac{1}{n} \sum (\log_{10}(L_{V(in)}) - \log_{10}(L_{V(in)}^{\gamma_{new}}))^2 \end{aligned} \quad (2)$$

where $loss(mse)$ is loss computed in terms of mean squared error (MSE)

In this context, γ is defined as a trainable parameter, initialized with an initial value of 1.0, and updated during training based on gradient of the loss function with respect to the gamma value ∇_{γ} , this allows automated adjustment of gamma value to a suitable level. The gamma is then updated based on (3). The pseudocode that explains how the gamma can be treated as trainable parameter and adjusted automatically during training is shown in Figures 2(a) and 2(b) (see in Appendix).

$$\gamma_{new} = \gamma_{old} + \nabla_{\gamma} \quad (3)$$

where γ_{new} is updated gamma value, γ_{old} is previous gamma value, and ∇_{γ} is gradient of a loss function with respect to the trainable parameter (gamma)

2.2. Human action recognition network

In this paper, the one-stage detector YOLOv7 [8] has been chosen as the HAR network, which is widely used in fast processing applications. Compared with other previous versions, YOLOv7 outperforms all in terms of speed with satisfied accuracy. Transfer learning has been utilized for the implementation of this HAR network based on training protocol in [8]. It refers to leverage of the knowledge gained by the pretrained YOLOv7's weight and feature representations from a large-scale coco dataset, then fine tuning the weight based on the new target data, which refers to human action in this case. This technique helps to save time and computational resources especially in the case of limited data as in this case. The same network architecture and loss functions as the original YOLOv7 have been adopted.

2.3. Data training

MIT-FiveK [14] dataset consisting of well-lit images has been chosen. This set of images are then augmented by adding several types of noises, then, the image pairs are used for training DFE. Figure 3 shows the augmented dataset.

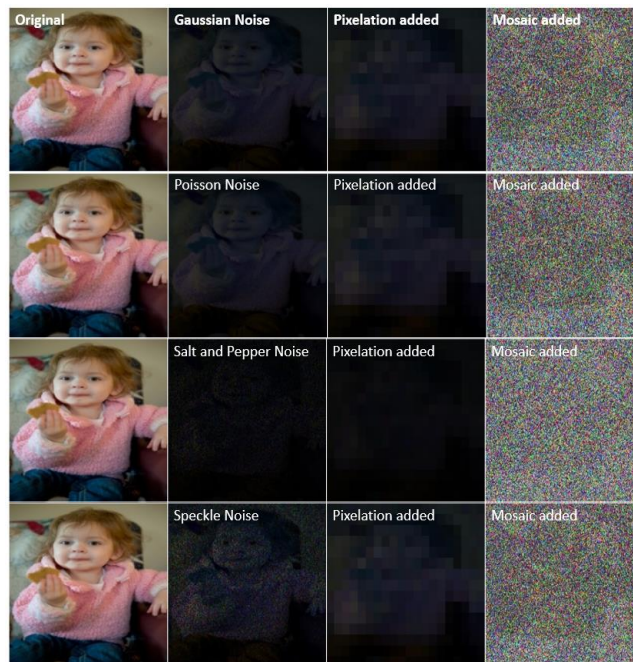


Figure 3. Dataset used for training DFE that consist of well-lit images and corresponding artificially generated dark images for DFE training

The purposed of DFE is to enhance the action recognition in dark (ARID) [15] video dataset. YOLOv7 has been trained on the enhanced version of ARID video frames with annotated bounding box coordinate and class label to achieve HAR. Videos in ARID dataset originally have frame rate of 30 FPS, many consecutive frames are nearly identical, hence contains insignificant and redundant information, so the frames have been reduced 10 times to just 3 FPS without changing the video duration.

MIT-FiveK dataset has total of 5,000 well-lit images, while ARID dataset consists total of 5,572 videos recording 11 classes of human actions (drink, jump, pick, pour, push, run, sit, stand, turn, walk, and wave) in poor lighting condition. To meet the hardware resource constraint and avoid overfitting, not all the MIT-FiveK images and ARID videos are selected for training. The configuration of train, validation and test data are made randomly and shown in Tables 1 and 2.

Table 1. Configuration of MIT-FiveK dataset

Train set	Validation set	Test set
Well-lit images and corresponding low light counterpart each own 3,000 images	Well-lit images and corresponding low light counterpart each own 750 images	400 artificially generated low light images

Table 2. Configuration of ARID dataset

Action classes	Train set	Validation set	Test set
Drink, jump, pick, pour, push, run, sit, stand, turn, walk, wave	Each action own 50	Each action own 6	Each action own 6
Total		628 clips	

Prior to training of the proposed model, data labeling was involved to provide region of interest (ROI), basically the human that performing action, to help algorithm identify elements of an image and return relevant result. The ROI in each video frame has been annotated with bounding boxes and action class labels using MATLAB video labeler with embedded point tracker algorithm to automate the labelling process. All the annotations have then been converted into YOLO compatible format.

3. RESULTS AND DISCUSSION

The effectiveness of the proposed method has been evaluated in dark scenario only. In this paper, the proposed DFE model has been trained with the Adam optimizer with a starting learning rate of 0.001 and batch size of 1 for 40 epochs. The training protocol of [8] has been adopted to perform transfer learning on YOLOv7 for HAR task. The video frames have been resized to 256×256 for training. The experiments use run on the 4GB RTX3050 and 16 GB Tesla T4 GPU.

3.1. Performance of dark frame enhancement

DFE has been trained under three conditions. These conditions are firstly, train without GIC technique (only the extracted good lighting features), secondly, train with additional fixed GIC technique (2.5 fixed gamma), and lastly, train with additional adaptive GIC technique (proposed model). The training and validation loss obtained from those three conditions are compared and illustrated in Figures 4(a) to 4(c) respectively. The DFE model with adaptive GIC shows smaller overall losses, where the losses exhibit small fluctuations from the start of epoch one until the last of epoch 40, as compared to the other two, this indicates that the proposed model able to enhance the darken images to be closer to the corresponding ground truth.

The computation of structural similarity index measure (SSIM) has been carried out on the test set. The SSIM is recorded in Figure 5. Image pairs enhanced by DFE without GIC and DFE with adaptive GIC technique have similar SSIM value of around 0.67, which show better enhancement performance compared to the one with fixed gamma GIC. Furthermore, the enhancement results are also compared in Figure 6. The frames enhanced by DFE with adaptive GIC technique are more visually receptive in terms of edges and color tones in the sense of human eyes and the optimum gamma is around 1.5 (shown in Figure 4(c)).

3.2. Performance of proposed method on human action recognition task in dark video

Three versions of enhanced ARID frames and the original dark frames (without enhancement) have been used for YOLOv7 training. The effect of different enhancement model's setting on the HAR model accuracy has been investigated in terms of Top-1 accuracy and is recorded in Figure 7. In terms of processing time for a single frame, the results are tabulated in Table 3. The proposed model (with adaptive GIC) always shows the longest processing time means the video frames enhanced contains more meaningful details that the proposed model can extract and process. Some of the recognition results on the dark video frames are shown

in Figure 8. It can be observed that when the DFE model with additional adaptive GIC technique is used for enhancing dark frames, the YOLOv7 model is able to predict the action more accurately than the other two. Nevertheless, some of the recognition results on the dark video frames are shown in Figure 8. It can be observed that when the DFE model with additional adaptive GIC technique is used for enhancing dark frames, the YOLOv7 model is able to predict the action more accurately than the other two.

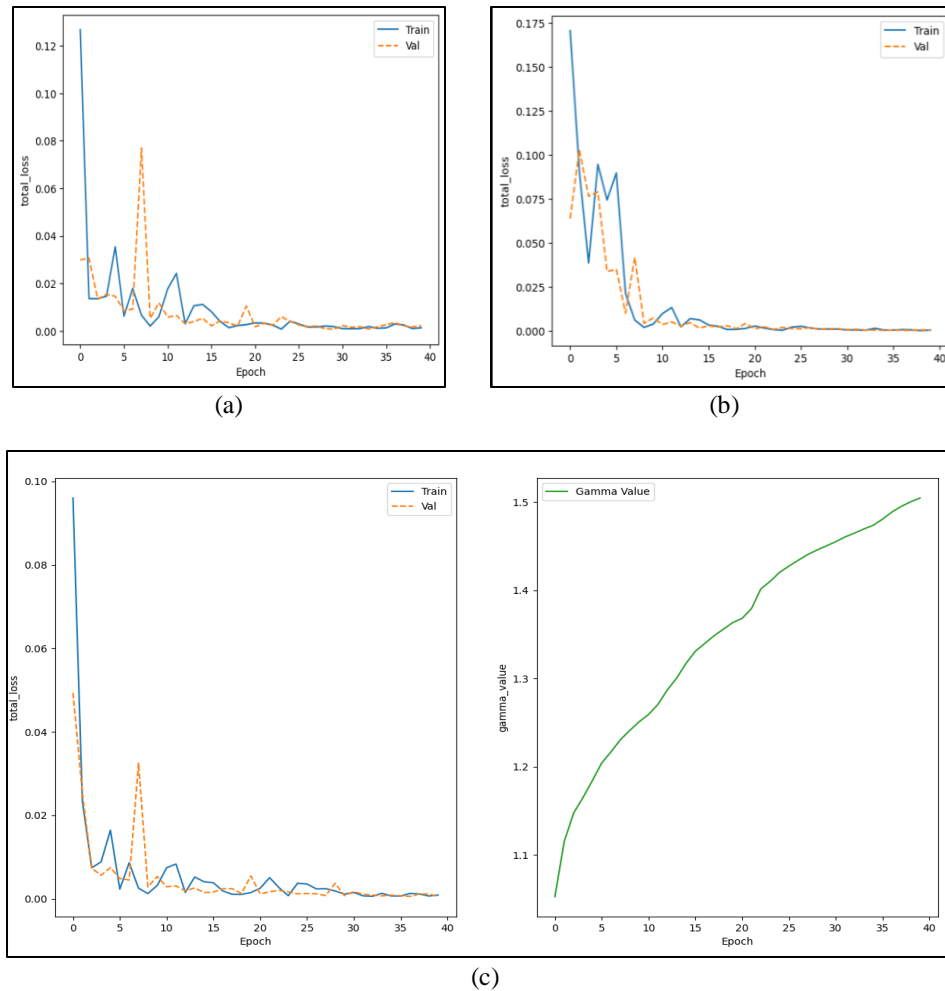


Figure 4. The proposed adaptive GIC technique contributes to a smaller loss: (a) loss when training without GIC, (b) loss when training with fixed gamma GIC, and (c) loss when training with adaptive GIC (default gamma value=1.0)

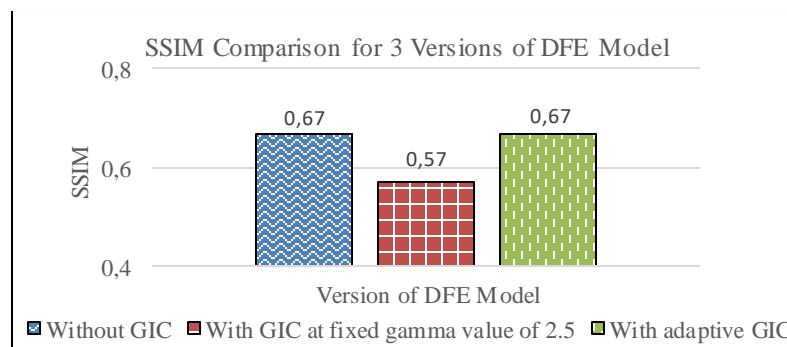


Figure 5. SSIM comparison for three versions of DFE model. The proposed adaptive GIC technique contributes to higher SSIM (better enhancement performance)

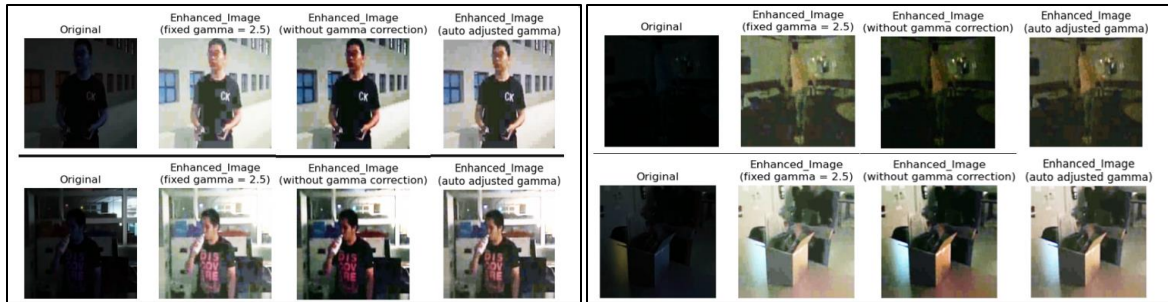


Figure 6. Dark frames from ARID dataset enhanced by DFE. The proposed adaptive GIC technique contributes to more visually receptive of enhancement results (shown by fourth column)

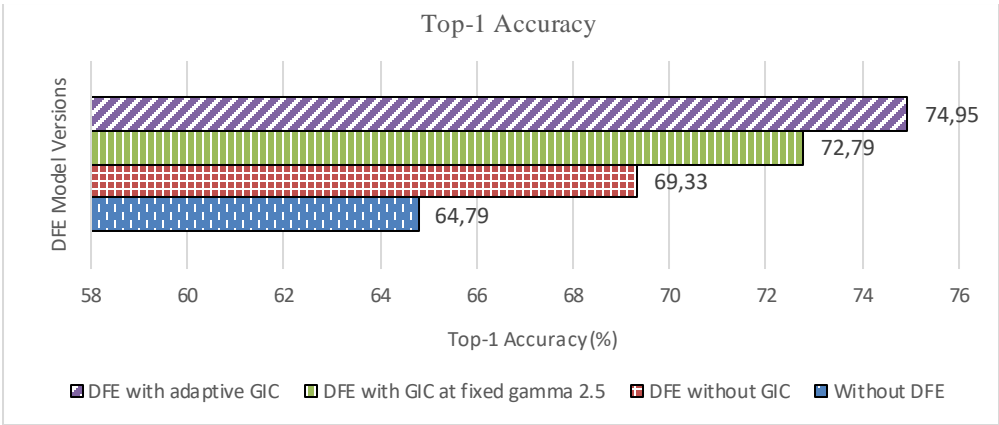


Figure 7. The proposed adaptive GIC technique contributes to highest Top-1 accuracy

Table 3. Single frame processing time comparison between different models			
Hardware accelerator setting	Enhancement setting	Single frame processing time (ms per frame)	Frame rate (fps)
4 GB RTX 3050 GPU	Without Enhancement	33.18	30
	DFE without GIC	89.99	11
	DFE + fixed gamma GIC	92.29	10
	DFE + adaptive GIC	93.99	10
	Without Enhancement	9.78	102
16 GB Tesla T4 GPU	DFE without GIC	16.40	60
	DFE + fixed gamma GIC	17.06	58
	DFE + adaptive GIC	17.59	56

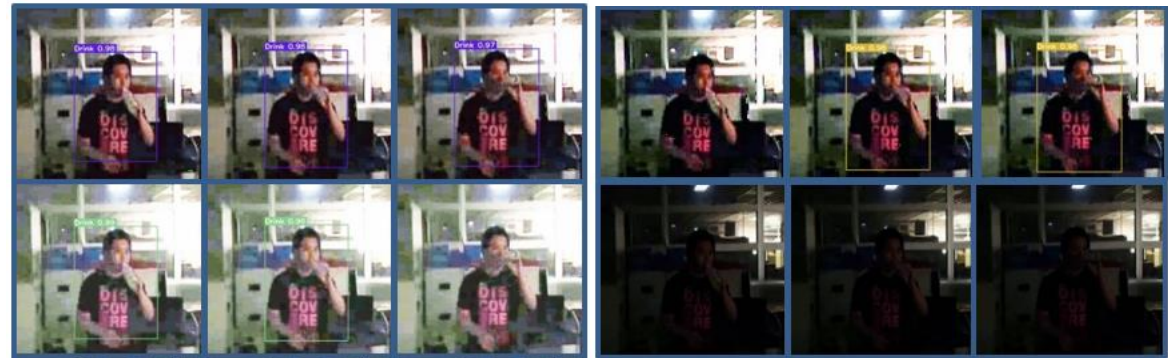


Figure 8. Detection results from YOLOv7 assisted by 3 DFE models. Upper-left: DFE with adaptive GIC. Bottom-left: DFE with fixed GIC. Upper-right: DFE without GIC, Bottom-right: Original frames

In short, Figure 8 shows that the Top-1 accuracy for YOLOv7-based HAR with DFE + adaptive GIC outperforms the one without any enhancement (64.79%) by 10.16%, the one without GIC (69.33%) by 5.62%, and the one with fixed GIC (72.79%) by 2.16%. It can be said that DFE model with additional adaptive GIC is the most effective algorithm among the others two to improve YOLOv7 accuracy for HAR. Therefore, the proposed model accuracy, base model size and total parameters have been further compared to benchmarked approaches as shown in Table 4. The proposed method has achieved good accuracy on par with previous research work and could be further improved by including temporal information to differentiate action with similar movement such as “stand” and “sit” in future work. The proposed model owns second smallest model size 74.8 MB and second highest total parameters 36.54 M indicate that it is having more capacity to learn and process more meaningful information from a frame, while still preserving the compatibility with application with limited hardware resources. This work focuses on the processing speed which overlook by previous research work, enabling more comprehensive analysis in the future.

Table 4. Comparison of top-1 accuracy and single frame processing time

Citation	Method	Image enhancer	Dataset	Top-1 accuracy	Processing time (ms/frame)	Base model size (mb)	Total parameters
[11]	DANorm + R(2+1)D-34	DANorm	ARID	80.73	-	487.8 MB [16]	-
[12]	Z-DEAF + R(2+1)D-34	Zero-DCE		49.39	-	487.8 MB [16]	-
[13]	C3D	GIC		39.17	-	-	34.8 M [17]
	3D-ShuffleNet			44.35	-	-	1.52M [18]
	3D-SqueezeNet			50.18	-	0.5 MB [18]	2.15 M [18]
	3D-ResNet-18			54.68	-	256.1 MB [16]	33.36 M [18]
	Pseudo-3D-199			71.39	-	98 MB [19]	-
	13D-Two-stream			73.39	-	-	25 M [20]
	3D-ResNext-101			74.73	-	-	48.75 M [18]
	Dark-Light + R(2+1)D-34			94.04	Computationally intensive [21]	487.8 MB [16]	-
	3D-ResNext-18			About 75.00	-	256.1 MB [16]	33.36 M [18]
[22]		GIC		About 77.00	-		
		KinD		About 69.00	-		
		HE+GIC		About 78.00	-		
				90.46	-		
[23]	Delta Sampling R(2+1)D-BERT	Zero-DCE		90.46	Computationally intensive 0	464 MB [24]	-
This work	DFE + Adaptive GIC + YOLOv7	DFE + Adaptive GIC		74.95	93.99 (RTX3050), 17.59 (Tesla T4) [24], [25]	74.8 MB [24]	36.54

4. CONCLUSION

A hybrid model for HAR in dark environments has been proposed. The model integrated DFE, adaptive GIC and YOLOv7 was proposed to improve ARID environments in terms of both accuracy and speed, where each of the dark video frames was adaptively enhanced to achieve better recognition performance. By taking advantage of adaptive GIC, the proposed approach was able to handle dark videos. The experimental results showed that the proposed method performed better than some previous approaches on the ARID dataset in terms of accuracy and outperformed the majority of the previous works in terms of base model size and total parameter.

ACKNOWLEDGEMENTS

The authors acknowledge the technical and financial support by the Ministry of Higher Education, Malaysia, under the research grant no. FRGS/1/2020/ICT02/UTEM/02/1 and Universiti Teknikal Malaysia Melaka (UTeM).

APPENDIX

```

# Declare initial gamma = 1.0
gamma = 1.0

# Define DFE model
Class DFE (imgSize = None):
    # Construction of CNN layers done here
    return DFE output

# Define loss function
def compute_loss (input_frame, DFE_output, gamma):
    # gamma correction applied here and get the MSE as loss
    enhanced_frame = DFE_output*(gamma)
    loss = sum of [(input_frame - enhanced_frame)2] for each frame / batch_size
    return loss

# Treat gamma as trainable parameter and update it during backpropagation.
def training_loop (input_frame):

    # get lighting feature as output of DFE model
    DFE_output = DFE (input_frame)

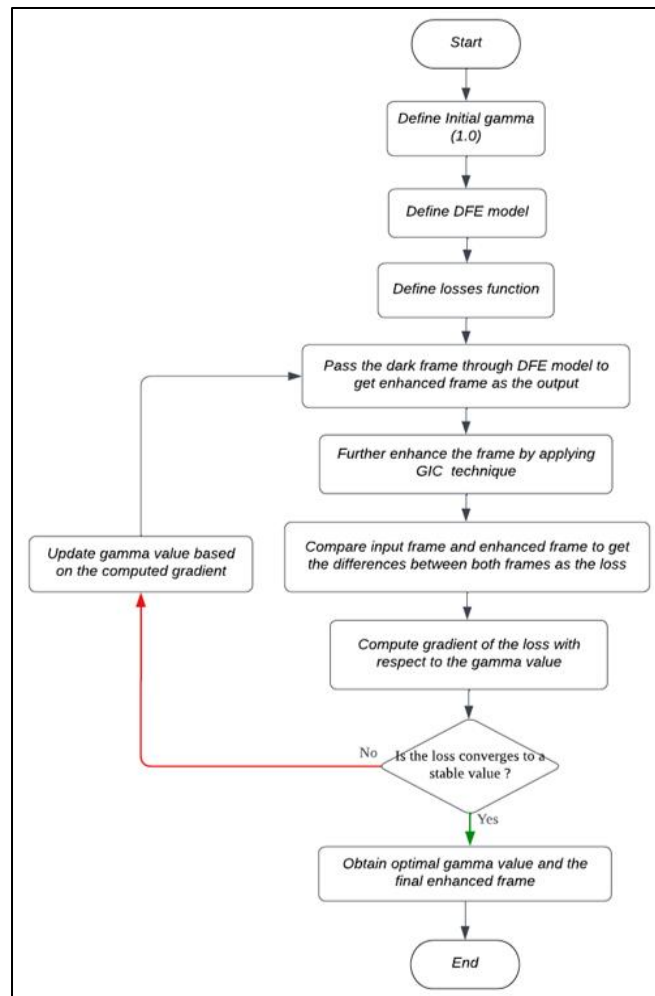
    # compute loss between input frame and enhanced frame
    loss = compute_loss (input_frame, DFE_output, gamma)

    # compute gradient of the loss with respect to trainable parameter
    gradients = tape.gradient(loss, DFE.trainable_weights + [gamma])

    # updated gamma value depending on gradient such that:
    gamma = gamma + gradient
    1. if loss increase at initial gamma = 1.0, gradient increase, gamma will increase to make dark image brighter.
    2. if loss is continue increase at latest gamma, gradient continue increase and gamma increase again.
    3. if loss now decrease at latest gamma, gradient will decrease resulting in decreasing gamma.
    4. these will repeat and repeat again until loss become steady at minimum, gradient approaches 0 and gamma
       changes unsignificantly, the optimal gamma is then obtained.

```

(a)



(b)




Figure 2. Updated gamma value: (a) pseudocode and (b) flowchart

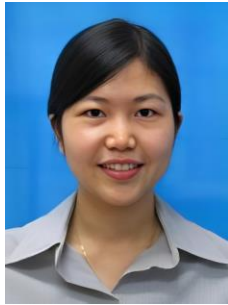
REFERENCES




- [1] C. Chen, Q. Chen, M. Do, and V. Koltun, "Seeing motion in the dark," in *Proceedings of the IEEE International Conference on Computer Vision*, IEEE, 2019, pp. 3184–3193, doi: 10.1109/ICCV.2019.00328.
- [2] H. H. Pham, L. Khoudour, A. Crouzil, P. Zegers, and S. A. Velastin, "Video-based human action recognition using deep learning: A review," *ArXiv-Computer Science*, pp. 1–25, Aug. 2022.
- [3] S. Ranasinghe, F. A. MacHot, and H. C. Mayr, "A review on applications of activity recognition systems with regard to performance and evaluation," *International Journal of Distributed Sensor Networks*, vol. 12, no. 8, pp. 1–22, 2016, doi: 10.1177/1550147716665520.
- [4] Y. Hu, H. He, C. Xu, B. Wang, and S. Lin, "Exposure: a white-box photo post-processing framework," *ArXiv-Computer Science*, pp. 1–23, Sep. 2017.
- [5] Y. Atoum, "Detecting objects under challenging illumination conditions," *P.h.D. Dissertation*, Department of Electrical Engineering, Michigan State University, Michigan, USA, 2018, doi: 10.25335/M5WK4H.
- [6] Y. Zhang, J. Zhang, and X. Guo, "Kindling the darkness: a practical low-light image enhancer," in *MM 2019 - Proceedings of the 27th ACM International Conference on Multimedia*, in MM '19. ACM, 2019, pp. 1632–1640, doi: 10.1145/3343031.3350926.
- [7] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: unified, real-time object detection," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2016, pp. 779–788, doi: 10.1109/CVPR.2016.91.
- [8] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, "YOLOv7: trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 7464–7475, doi: 10.1109/cvpr52729.2023.00721.
- [9] S. Shinde, A. Kothari, and V. Gupta, "YOLO based human action recognition and localization," *Procedia Computer Science*, vol. 133, pp. 831–838, 2018, doi: 10.1016/j.procs.2018.07.112.
- [10] Z. Y. Li, S. Y. Lin, Z. M. Liang, Y. J. Lei, Z. F. Wang, H. Chen, "PDE: A real-time object detection and enhancing model under low visibility conditions," in *2022 International Journal of Advanced Computer Science and Applications*, vol. 13, no. 12, doi: 10.14569/IJACSA.2022.0131299.
- [11] Z. Liang et al., "Domain adaptable normalization for semi-supervised action recognition in the dark," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, 2022, pp. 4250–4257, doi: 10.1109/CVPRW56347.2022.00470.
- [12] Z. Chen, Z. Fan, Y. Li, H. Gao, and S. Lin, "Z-domain entropy adaptable flex for semi-supervised action recognition in the dark," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, IEEE, 2022, pp. 4258–4265, doi: 10.1109/CVPRW56347.2022.00471.
- [13] R. Chen, J. Chen, Z. Liang, H. Gao, and S. Lin, "Darklight networks for action recognition in the dark," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, 2021, pp. 846–852, doi: 10.1109/CVPRW53098.2021.00094.
- [14] V. Bychkovsky, S. Paris, E. Chan, and F. Durand, "Learning photographic global tonal adjustment with a database of input/output image pairs," in *IEEE Computer Society conference on Computer Vision and Pattern Recognition*, IEEE, 2011, pp. 97–104, doi: 10.1109/cvpr.2011.5995332.
- [15] Y. Xu, J. Yang, H. Cao, K. Mao, J. Yin, and S. See, "ARID: a new dataset for recognizing action in the dark," in *Communications in Computer and Information Science*, Springer Singapore, 2021, pp. 70–84, doi: 10.1007/978-981-16-0575-8_6.
- [16] K. Hara, H. Kataoka, and Y. Satoh, "Can spatiotemporal 3D CNNs retrace the history of 2D CNNs and ImageNet?," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6546–6555, doi: 10.1109/CVPR.2018.00685.
- [17] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3D convolutional networks," in *Proceedings of the IEEE International Conference on Computer Vision*, IEEE, 2015, pp. 4489–4497, doi: 10.1109/ICCV.2015.510.
- [18] O. Kopuklu, N. Kose, A. Gunduz, and G. Rigoll, "Resource efficient 3D convolutional neural networks," in *Proceedings-2019 International Conference on Computer Vision Workshop, ICCVW 2019*, IEEE, 2019, pp. 1910–1919, doi: 10.1109/ICCVW.2019.00240.
- [19] Z. Qiu, T. Yao, and T. Mei, "Learning spatio-temporal representation with pseudo-3D residual networks," in *Proceedings of the IEEE International Conference on Computer Vision*, IEEE, 2017, pp. 5534–5542, doi: 10.1109/ICCV.2017.590.
- [20] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *Proceedings-30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, IEEE, 2017, pp. 4724–4733, doi: 10.1109/CVPR.2017.502.
- [21] A. Ulhaq, "Adversarial domain adaptation for action recognition around the clock," *2022 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, pp. 1–6, 2022, doi: 10.1109/DICTA56598.2022.10034580.
- [22] H. R. Patel and J. T. Doshi, "Human action recognition in dark videos," *IEEE International Conference on Artificial Intelligence and Machine Vision*, pp. 1–5, 2021, doi: 10.1109/AIMV53313.2021.9670923.
- [23] S. Hira, R. Das, A. Modi and D. Pakhomov, "Delta sampling R-BERT for limited data and low-light action recognition," *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 853–862, 2021, doi: 10.1109/CVPRW53098.2021.00095.
- [24] A. Mahmoudi, O. Amel, S. Stassin, M. Liagre, M. Benkedadra, and M. Mancas, "A review and comparative study of explainable deep learning models applied on action recognition in real time," *Electronic*, vol. 12, no. 9, May 2023, doi: 10.3390/electronics12092027.
- [25] O. Sharir, B. Peleg, and Y. Shoham, "The cost of training NLP models: a concise overview," *ArXiv-Computer Science*, pp. 1–6, 2020.

BIOGRAPHIES OF AUTHORS






Shi Yong Goh    recently graduated with a bachelor's degree in electronic engineering from Universiti Teknikal Malaysia Melaka (UTeM). He conducted his final year project in the system engineering field, which involved development of a computer vision algorithm. Throughout his academic journey, he engaged in extracurricular activities, such as Institute of Engineers, Malaysia (IEM), Institute of Electrical and Electronics Engineers (IEEE) and Google Developer Student Club (GDSC). Now a graduate, he is eager to launch his career in R&D, system integration, IC design and other related fields. He can be contacted at email: shiyong0606@gmail.com






Yan Chiew Wong    is an Associate Professor at Faculty of Electronics and Computer Technology and Engineering, Universiti Teknikal Malaysia Melaka (UTeM). She completed her doctorate from The University of Edinburgh, United Kingdom in 2014. She has more than nine years of industry experience in semiconductor design. She has previously worked at Infineon Technologies, Intel and Sofant Technologies. She has been involved intensively in the research of IC design, applied intelligent computing, and power management/harvesting design. She is a recipient of different national and international awards such as ITEX, INNOVATE, IEM, and EDS societies. She equipped with engineering design and implementation skills. She can be contacted at email: ycwong@utem.edu.my.



Syafeeza Ahmad Radzi    earned her Bachelor of Engineering degree in Electrical-Electronic Engineering in 2003 and subsequently completed her Master's degree in Electrical – Electronic & Telecommunication Engineering in 2005, both at Universiti Teknologi Malaysia. She attained her Doctor of Philosophy (Ph.D.) in Electrical Engineering from the same institution in 2014. Currently, she holds the position of Associate Professor at the Faculty of Electronics and Computer Technology and Engineering, Universiti Teknikal Malaysia Melaka (UTeM). Her extensive research portfolio encompasses diverse areas, including embedded systems, pattern recognition, machine learning, image processing, and biometrics. She can be contacted at email: syafeeza@utem.edu.my.



Ranjit Singh Sarban Singh    is an Associate Professor at the School of Engineering and Technology, Sunway University, Malaysia. He completed his doctorate from Brunel University, London United Kingdom in 2016. He has 3 years of industrial working experience as an engineer with Western Digital Sdn. Bhd. As an academician his research area mostly focused on embedded system design – battery management system, renewable energy system development, image processing. He has also won numerous national and international award at ITEX, MTE, and INNOVA Brussels. He is also actively engaging with industries to equip himself with industrial knowledge, which can be shared during his teaching and learning activities with students. He can be contacted at email: ranjits@sunway.edu.my.